Investigating the use of recurrent motion modelling for speech gesture generation

Ylva Ferstl*
Graphics, Vision and Visualisation Group
Trinity College Dublin
yferstl@tcd.ie

Rachel McDonnell
Graphics, Vision and Visualisation Group
Trinity College Dublin
ramcdonn@tcd.ie

ABSTRACT

The growing use of virtual humans demands generating increasingly realistic behavior for them while minimizing cost and time. Gestures are a key ingredient for realistic and engaging virtual agents and consequently automatized gesture generation has been a popular area of research. So far, good gesture generation has relied on explicit formulation of if-then rules and probabilistic modelling of annotated features. Machine learning approaches have yielded only marginal success, indicating a high complexity of the speech-to-motion learning task. In this work, we explore the use of transfer learning using previous motion modelling research to improve learning outcomes for gesture generation from speech. We use a recurrent network with an encoder-decoder structure that takes in prosodic speech features and generates a short sequence of gesture motion. We pre-train the network with a motion modelling task. We recorded a large multimodal database of conversational speech for the purpose of this work.

CCS CONCEPTS

• **Computing methodologies** → **Animation**; *Machine learning*;

KEYWORDS

character animation, motion synthesis, behavior generation, recurrent networks, deep learning

ACM Reference Format:
Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of IVA '18: International Conference on Intelligent Virtual Agents (IVA '18).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3267851.3267898

1 INTRODUCTION

Virtual humans are becoming more and more popular for many applications, such as video games, human-computer interfaces (e.g., virtual museum guides [35]), virtual reality entertainment, and personalized training (e.g., virtual patients for medical training [19]), including training of interpersonal skills. Interaction with virtual agents is more engaging when their verbal output is coupled with appropriate gesturing behavior [33], and realistic gestures are essential for adequately mimicking real human interactions, in which non-verbal behaviour plays a major role in conveying information [16, 32]. Furthermore, users can detect whether a virtual human's gestures are consistent with the produced speech [13].

Classic approaches to animating conversing virtual humans consist of specifically recording motion or hand-crafting animations for predefined utterances (e.g., [11, 18]), which can be both time-consuming and expensive. As an alternative approach, research has explored methods to automatically generate animation for virtual humans from speech. Speech and speech-accompanying motion is said to arise from the same internal process and reflect the same semantic meaning [5, 26], suggesting the theoretical potential to infer one from the other to some extent. Inferring gesture motion directly from speech has however proven to be a very complex learning task with limited success.

In our work, we apply research of human motion modelling and define the speech-to-gesture learning objective as a transfer learning task. Namely, we pre-train part of our network with a state-of-the-art human motion modelling approach before training the final speech-to-gesture model. We compare two recurrent network architectures.

Additionally, we recorded a multimodal dataset of conversational speech, containing 4 hours of audio, motion, and video data from one actor.

2 RELATED WORK

The research into automatic motion generation for speech can largely be divided into three categories, namely rule-based systems, statistical modelling methods, and machine learning approaches.

Rule-based systems require the explicit formulation of phrase-to-gesture rules and can generate expressive gestures. The advantage of rule-based systems is the possibility to capture semantic content, allowing for meaningful and even rhetorical gestures (examples include [6, 24, 37]). However, their expressiveness is directly limited by the number of rules created, with the gesture behavior based on the ideas of the rule creator rather than actual human gesturing behavior. A more sophisticated approach uses cognitive models that rely on building a multimodal memory representing visual-spatial properties of phrases, another time-intensive approach that allows for semantic coordination of speech and gesture [1].

Statistical modelling methods use estimated conditional probabilities of certain speech features co-occuring with specific motion features. For example, in a work by Neff et al. [29], a video corpus is annotated using approximately 90 semantic tags for the speech channel, and 30 gesture tags, including information such as which

*Corresponding author

IVA '18, November 5–8, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6013-5/18/11…$15.00
https://doi.org/10.1145/3267851.3267898

spoken word is associated with the tagged gesture. With this, a statistical model is computed that can subsequently take in new, annotated text representing the speech, and generate meaningful, naturalistic gestures. In a similar approach, Bergmann and Kopp [2, 3] have used Baysian decision networks.

Machine learning approaches for gestures can learn from motion data produced by conversing humans and, like statistical models, do not rely on the explicit formulation of if-then rules. These approaches have largely focused on the generation of so-called beat gestures, simple, repetitive motions that accompany speech rhythm. The kinematics of beat gestures (e.g., speed and acceleration) have been shown to correlate with the prosodic features of speech. Prosodic speech features can be extracted quickly and easily, and contain information such as emphasis and emotion. Beat gesture systems rely purely on analysis of such prosodic features from speech without extraction of semantic content [4, 7, 8, 21, 22]. A variety of machine learning models have been used in this domain, including hidden Markov models and conditional random fields. Related works have used similar methods for generating head motion, which can be seen as a similar type of speech motion as it is comprised of simple beat-like movements. In this line of work, the yaw, pitch and roll of the head are modelled based on prosodic speech features. Good results for head motion generation have been obtained with deep neural networks [12], and more recently bidirectional recurrent network architectures [17].

Going beyond beat gestures, Chiu and Marsella [9] have published work that models more complex gestures using deep learning without explicit rules. The authors combine a deep neural network with conditional random fields to model the temporal dynamics of gesture motion. This method showed improvement compared to previous work on gesture prediction. However, the authors rely on predicting a small set of predefined gestural signs, limiting the gesture output to this set. Furthermore, gestural signs needed to be hand-annotated in the training set, limiting the amount of training data that can be feasibly acquired. A recent work by Takeuchi et al. [36] also used a network architecture with temporal modelling capacity, generating gesture motion from Japanese speech. Their architecture is similar to the head motion model of Greenwood et al. [17], using a bidirectional recurrent network to generate motion directly from speech, however with only marginal success.

Gesture generation has been shown by this previous research to be a very complex problem and we may need to model the underlying speech and gesture processes for good results. In this work, we investigate whether modelling of the gesture motion may improve the speech-to-gesture prediction abilities of a network.

We explored the potential benefit of exploiting the extensive research done in the area of human motion modelling for improving the speech-to-motion modelling objective. We considered gesture generation from speech as a transfer learning task where a model is pre-trained with a motion-to-motion task before training the final speech-to-motion task. Transfer learning allows knowledge gained from training one problem to be applied to a related problem. For example, an image classifier trained on a large publicly available image dataset could be re-utilized for learning to label objects in a smaller, domain specific dataset. Transfer learning has shown success in a variety of domains, such as machine translation [38], image classification [31], and visual emotion recognition [30]. In this work, our pre-training task is predicting the next few frames of a motion sequence, receiving as an input the preceding motion sequence. This requires a modelling of the dynamics of human motion.

State-of-the-art work has shown the potential of recurrent neural networks for modelling human motion [15, 20, 25]. Recurrent networks can model sequential data by using recurrent connections between network activations at consecutive timesteps. For human motion modelling, recurrent networks seem to be able to capture the dynamics of a motion pattern well. Here, we apply the motion model proposed by Martinez et al. [25], which yielded good results with a relatively simple and fast-training architecture. We hoped that applying previous knowledge about the motion domain could decrease the complexity of the notoriously hard to model speech-to-motion relationship.

Recurrent networks have furthermore shown outstanding success in language modeling (e.g., [27]), specifically tasks such as machine translation (e.g., [23, 34]), which takes a sequence of text in language A and generates the same sequence for language B. We can consider the task of generating gestures from speech as a kind of machine translation where the audio signal is language A, and the synchronous motion is language B.

3 SPEECH AND GESTURE DATABASE

Previous research into speech gesture generation has used small datasets [7, 8], non-English datasets [36], or hand-annotated video data [9, 28]. No large, freely available corpus of multi-modal recordings of English conversational speech was known to us and we hence recorded our own dataset for the purpose of modelling the speech-motion relationship. The dataset consists of 244 minutes of high-quality motion capture, audio, and video data.

3.1 Data collection

We invited a single male actor for multiple recording sessions. The actor is a male native English speaker producing spontaneous and natural conversational speech without interruptions, i.e., without verbal cues from a conversation partner. The actor was instructed to speak freely and spontaneously about any topic he chose, including hobbies, daily activities, and movies. The actor speaks in a colloquial manner with a happy disposition and includes a large quantity of gesture motions.

The actor was addressing a person situated behind the camera in order to give him the visual feedback of a conversation partner. Each recording take was approximately ten minutes long. We captured 23 takes, totalling 244 minutes of data. Two additional takes of eight minutes each are available of video and audio data, without motion capture.

The actor's motion was captured with a 53 marker setup and 20 Vicon cameras at 59.94 Frames per Second (FPS). Audio was recorded at 44 kHz. Video was captured with a single HD camera placed between the actor and the person he was addressing.

3.2 Data processing

We experimented with different representations of the motion for our learning task, such as Euler angles and quaternions, but finally used the raw joint angles in exponential map format, as proposed

Figure 1: View of the camera in the multimodal conversational speech database. The actor is always addressing a listener situated right behind the camera. The actor was allowed to move freely with the restriction to stay in good view of the camera.

by Fragkiadaki et al. [15] and used by Martinez et al. [25]. Two takes were selected as validation data, representing about 8% of the total data. During each validation step, 8 seeds are randomly selected from the validation set to compute the validation loss.

We explored different audio features, but in the final version of this work used the log of the 27 values of the Mel-scaled spectrogram with no cosine transforms, computed from FFT magnitude. The more primitive mel-frequency filter bank values have outperformed MFCC features in previous works in both pure speech modelling [10] and speech to motion modelling [12]. We extracted the audio features with openSMILE [14].

4 RECURRENT SPEECH-MOTION MODEL

We experimented with two variations of a sequence-to-sequence architecture for learning the speech to motion prediction task. In the first setup, we predict a motion sequence directly from a speech sequence, with both modalities sharing one embedding between encoder and decoder. In the second setup, we insert an additional recurrent layer between speech encoder and motion decoder. Our architectures are based on the motion modelling network proposed by Martinez et al. [25]. We downsample our data to 50 frames per second.

4.1 Speech to motion

Both encoder and decoder consist of a single recurrent cell (a gated recurrent unit (GRU)) of size 1024 with residual connections between the output and input of each cell. The residual connections are skip connections that allow a copy of the input to skip the encoder cell to directly feed into the output of the cell. The loss is computed as the Euclidean distance between the predicted motion sequence and the ground truth in angle space.

We first pre-train the network on a motion prediction task with residual connections modelling motion velocity, as described in Martinez et al. [25]. However, we do not tie weights between encoder and decoder, allowing encoder and decoder weights to be updated separately. We train this sequence-to-sequence architecture to predict the next 15 motion frames based on an input motion sequence of 200 frames. We use our complete set of motion data for this training task. We train the network for 11,000 iterations, results are visualized in Figure 2. We then take the learned weights of the decoder and reuse for the final speech-to-motion training task.

In the speech-to-motion training task, we predict 20 motion frames based on both the corresponding 20 audio frames and 180 preceding audio frames. Hence at each prediction step, the network gets a context of 4 seconds, and predicts the final 0.4 seconds of gesture motion. We empirically found prediction of longer motion sequences hard to learn, but outputs of multiple overlapping speech sequences could theoretically be concatenated into longer motion sequences.

We run the same network without motion pre-training to evaluate the benefits of transferring the motion model knowledge to the speech to motion models.

4.2 Deep speech to motion

In our second architecture, we added an additional recurrent layer as connection between audio encoding and a motion decoding. The additional layer consists of a single recurrent cell (also a GRU) of size 1024, also with residual connections. The output state of this additional cell is passed to a motion decoder that was pre-trained on a motion modelling task as described in the previous section.

We again compare the results of the pre-trained versus not pre-trained model.

4.3 Results

We first run pre-training for an initial 10 thousand iterations as suggested by the model's authors [25], we then continue training for a thousand iterations at a time to check for possible improvements. The results of this are plotted in Figure 2. We train the model for 100 thousand iterations, after which training has slowed down and significant further improvement would take an unreasonable amount of time. We use the resulting model weights for initializing the decoder of our speech to motion models.

We train both of our architectures for 100 thousand iterations. For the deep speech to motion architecture, we first experimented with fixing the weights of the motion decoder and only trained the preceding layers. This was to try and focus on learning a motion representation from speech. However, this did not result in a good learning trajectory and we hence let the model update all weights during training. Though loss and validation loss still appeared to be dropping for both models after 100 thousand iterations, training has stagnated at that point. Both networks reached a minimum of approximately 0.38 in angle loss error at that point. The evolution of the training errors is shown in Figure 3.

We compare our pre-trained network's performances to the same architectures without motion pre-training. Surprisingly, as we can see in Figure 4, pre-training only appears to help at the beginning of

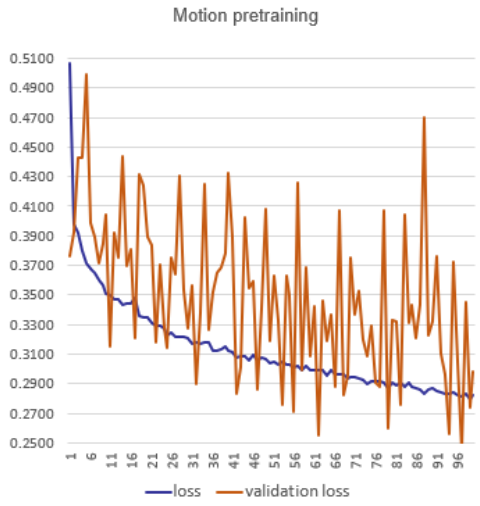

Figure 2: Results of motion pre-training. Plotted is the mean angle loss over one evaluation period, and the validation loss at the evaluation step. One evaluation period consists of $x * 10^3$ iterations. We train the model for 100 thousand iterations.

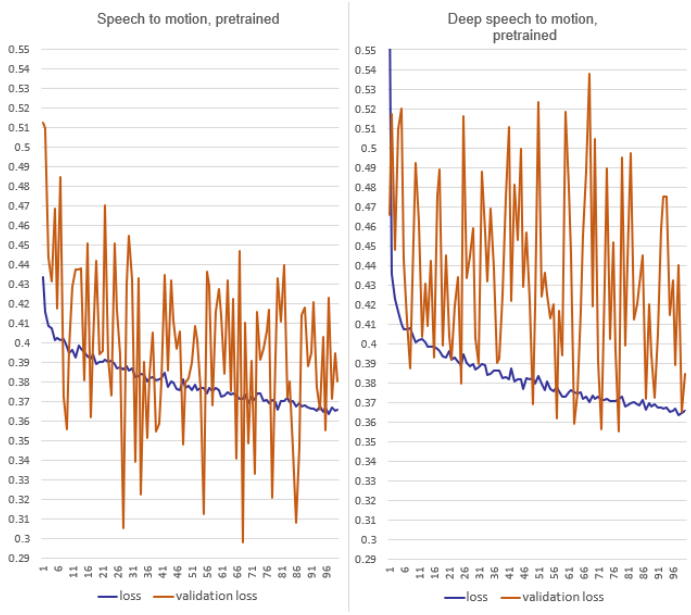

Figure 3: Training results of the speech to motion models after pre-training with motion modelling. Plotted is the mean angle loss over one evaluation period, and the validation loss at the evaluation step. One evaluation period consists of $x * 10^3$ iterations.

the training, essentially speeding up convergence, before stagnating at approximately the same error value.

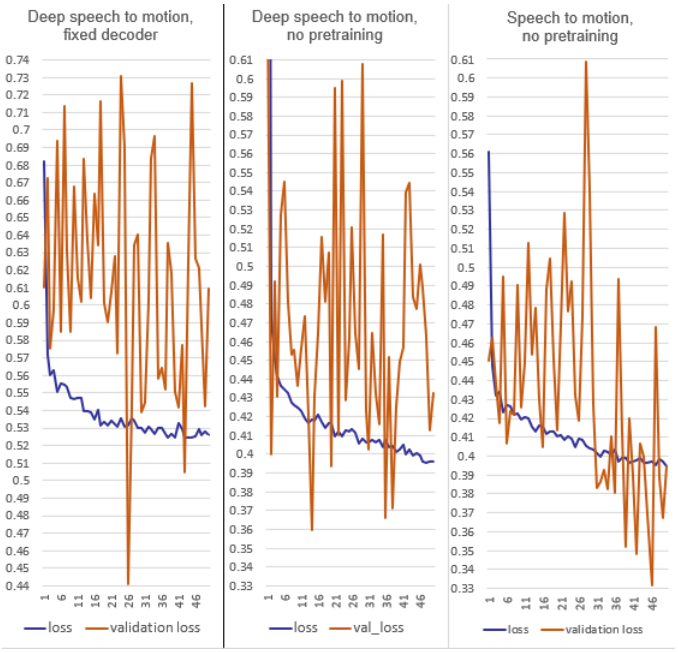

Figure 4: *Left*: Results of training the deep speech to motion model with fixed decoder weights. *Right*: Results of training both speech to motion models without prior motion modelling. Plotted is the mean angle loss over one evaluation period, and the validation loss at the evaluation step. One evaluation period consists of $x * 10^3$ iterations.

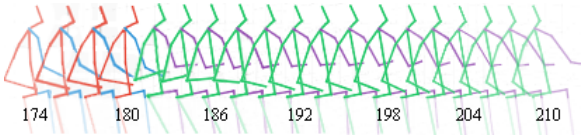

Figure 5: Example of a predicted motion sequence from the speech-to-motion model. Annotated is the respective frame, prediction starts at frame 180.

5 DISCUSSION

Our models reach a relatively low error for short-term motion prediction from speech, and we do not observe a significant difference in performance between our two network architectures. Both the speech to motion, and the deep speech to motion network converge to similar error results. Notably, for the deep network, we were forced to open the decoder weights to further updating during training, as the pre-trained weights did not yield good results. The poor performance of the pre-trained motion decoder becomes further apparent when comparing network performance with and without prior motion modelling. We found no large advantage of training our networks on a pure motion modelling task before attempting the speech to motion prediction task.

It is interesting that transfer learning of a motion model did not help our recurrent network's performance much. It is possible that the pre-training did not actually learn a good enough motion

embedding. This would be supported by the finding that plugging the pre-trained model's decoder into our deep model and only training the encoding and connection layer did not yield good results. Martinez et al. [25] do note in their work that aperiodic motion such as during a discussion (as opposed to periodic motion such as walking), remain hard to model in general. The dynamics of gesture motion might be better understood by focusing on the actual speech flow, as opposed to modelling an inherent motion representation (i.e., gestural motion may not be well suited for modelling without the context of speech flow).

The work of Martinez et al. [25] also suggests that the motion model may be improved by jointly learning on multiple different actions. We only trained our model on gesture motion and hence incorporating data from multi-action databases might improve the outcome.

It is possible that the speech to motion learning task is not as related to the motion-to-motion task as we had assumed. The speech to motion 'translation' is arguably more complex than motion forecasting and might require much more than a rudimentary understanding of human motion dynamics.

We succeeded to some extent in generating short term motion generation and the output of multiple overlapping speech input sequences could theoretically build a longer term gesture sequence. However, we have yet to test how this method of motion generation holds up over longer speech sequences, and how the predicted motion is perceived by human observers. The results for short term prediction indicate that gesture generation directly from speech does work to some extent; a basis upon which more sophisticated future versions of the model can be built. In our work, we do not distinguish between generation of beat gestures and semantically meaningful gestures. However, production of iconic (e.g., outlining a round shape when describing a round object), symbolic (e.g., a thumbs up), and deictic (e.g., pointing) gestures would arguably require more elaborate modelling or a large dataset rich in these gestures.

We intend to share our large multimodal database with the research community. Our dataset could be useful not only for human motion research, but is also applicable for video and image recognition, as well as speech research.

6 FUTURE WORK

In future work, we would like to investigate the potential benefit of language modelling over motion modelling. We may view the speech to motion learning task as more of a translation between two, if very different, languages, rather than a motion modelling problem. Exploring the use of state-of-the-art speech recognition or translation models that have been trained on very large datasets could help reduce the complexity of modelling the speech to motion relationship.

While the motion model itself might benefit from a larger variety of actions to model as previously mentioned, our speech to motion model might also improve with a larger variety of conversational data. We currently only have data from one actor and it is difficult to measure how well this specific person's gestures can be modelled. Further research is needed for devising the 'perfect' dataset for speech gesture modelling and combining of researchers' resources may be helpful.

7 CONCLUSION

In conclusion, we do not find a significant benefit of applying transfer learning from a motion model. We showed some success in short term gesture generation from speech features, and would like to build upon these findings by better understanding the factors that make up the complexity of the task.

We contribute a multimodal database of conversational speech, containing about 4 hours of audio, motion, and video data of a single actor.

ACKNOWLEDGMENTS

This research was funded by Science Foundation Ireland under the ADAPT Centre for Digital Content Technology (Grant 13/RC/2016).

REFERENCES

[1] Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp. 2013. Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In *International Workshop on Intelligent Virtual Agents*. Springer, 203–216.

[2] Kirsten Bergmann and Stefan Kopp. 2009. GNetIc–Using bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents*. Springer, 76–89.

[3] Kirsten Bergmann and Stefan Kopp. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 361–368.

[4] Elif Bozkurt, Yücel Yemez, and Engin Erzin. 2016. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication* 85 (2016), 29–42. https://doi.org/10.1016/j.specom.2016.10.004

[5] Justine Cassell, David McNeill, and Karl-Erik McCullough. 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition* 7, 1 (1999), 1–34.

[6] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: the Behavior Expression Animation Toolkit. *ACM Transactions on Graphics* (2001), 477–486. https://doi.org/10.1007/978-3-662-08373-4_8

[7] Chung Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6895 LNAI (2011), 127–140. https://doi.org/10.1007/978-3-642-23974-8_14

[8] Chung-cheng Chiu and Stacy Marsella. 2014. Gesture Generation with Low-Dimensional Embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 781–788.

[9] Chung-Chen Chiu and Stacy Marsella. 2015. Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach. *International Conference on Intelligent Virtual Agents* 9238 of th, August 2015 (2015), 152–166. https://doi.org/10.1007/978-3-319-21996-7_17

[10] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. 2013. Recent advances in deep learning for speech research at Microsoft. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 8604–8608.

[11] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[12] Chuang Ding, Lei Xie, and Pengcheng Zhu. 2014. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications* 74, 22 (2014), 9871–9888. https://doi.org/10.1007/s11042-014-2156-2

[13] C Ennis, R McDonnell, and C O'Sullivan. 2010. Seeing is believing. *ACM Transactions on Graphics* 29 (2010), 91. https://doi.org/10.1145/1833351.1778828

[14] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[15] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Computer Vision (ICCV), 2015*

IEEE International Conference on. IEEE, 4346–4354.

[16] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in Cognitive Sciences* 3, 11 (1999), 419–429. https://doi.org/10.1016/S1364-6613(99)01397-2

[17] David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting head pose in dyadic conversation. In *International Conference on Intelligent Virtual Agents*. Springer, 160–169.

[18] Arno Hartholt (project leader). 2007-present. Gunslinger. http://ict.usc.edu/prototypes/gunslinger/

[19] Robert C. Hubal, Paul N. Kizakevich, Curry I. Guinn, Kevin D. Merino, and Suzanne L. West. 2000. The virtual standardized patient. *Medicine Meets Virtual Reality* (2000), 133–138. https://doi.org/10.3233/978-1-60750-914-1-133

[20] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-RNN: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5308–5317.

[21] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. *ACM Transactions on Graphics* 29, 4 (2010). https://doi.org/10.1145/1833351.1778861

[22] Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics* 28, 5 (2009), 1. https://doi.org/10.1145/1618452.1618518

[23] Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. 2014. A recursive recurrent neural network for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1491–1500.

[24] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 25–35. https://doi.org/10.1145/2485895.2485900

[25] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4674–4683.

[26] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[27] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

[28] Izidor Mlakar, Zdravko Kacic, and Matej Rojc. 2013. TTS-driven synthetic behaviour-generation model for artificial bodies. *International Journal of Advanced Robotic Systems* 10 (2013), 1–20. https://doi.org/10.5772/56870

[29] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 5.

[30] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 443–449.

[31] Long D Nguyen, Dongyun Lin, Zhiping Lin, and Jiuwen Cao. 2018. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on*. IEEE, 1–5.

[32] Regina Pally. 2008. A Primary Role for Nonverbal Communication in Psychoanalysis. *Psychoanalytic Inquiry* 21, 1 (2008), 71–93. https://doi.org/10.1080/07351692109348924

[33] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (2011), 247–252. https://doi.org/10.1109/ROMAN.2011.6005285

[34] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[35] William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, Chad Lane, Jacquelyn Morie, Priti Aggarwal, Matt Liewer, Jen Yuan Chiang, Jillian Gerten, Selina Chu, and Kyle White. 2010. Ada and grace: Toward realistic and engaging virtual museum guides. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6356 LNAI (2010), 286–300. https://doi.org/10.1007/978-3-642-15892-6_30

[36] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-Gesture Generation: A Challenge in Deep Learning Approach with Bi-Directional LSTM. *Proceedings of the 5th International Conference on Human Agent Interaction* (2017), 365–369. https://doi.org/10.1145/3125739.3132594

[37] Marcus Thiebaux, Stacy Marsella, Andrew N Marshall, and Marcelo Kallmann. 2008. SmartBody: behavior realization for embodied conversational agents. In *Proceedings of International Joint Conference on Autonomous Agents and Multiagent Systems*. 151–158. https://doi.org/10.1016/j.ins.2009.01.020

[38] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201* (2016).